# Weather and Crops: An Exploration of Agrotech

April 28, 2022

| | |
|---|---|
| Registration number: | 2100374 |
| Project: | Agrotech |
| Link to GitHub: | https://github.com/rigovm101/CE888_DataScience/tree/main/Project |

| | |
|---|---|
| Executive summary (max. 250 words) | 73 |
| Introduction (max. 600 words) | 273 |
| Data (max. 300 words/dataset) | 249 |
| Methodology (max. 600 words) | 113 |
| Results and Discussion (max. 1000 words combined) | 200 |
| Conclusions (max. 500 words) | 101 |
| Total word count | 1,009 |

## Contents

**Abstract**

This project aims to build a Machine Learning model capable of using the Agrotech dataset to predict the growth of crops using historic weather information, as well as information about the planting process and the crops themselves. The *ScikitLearn* Python Library was used to develop and test four Machine Learning models, with the best model achieving a $R^2$ score of 0.82, with a RMSE score of 30.73 on the prediction of its three labels.

# 1 Introduction

Waste exists in a lot industries. With the increase on population and changes in consumption habits, companies have had to adapt to these changes and cut down on waste. Another rapidly changing factor in todays world is weather. Climate change has brought a lot of uncertainty to our daily activities. An industry which is at the intersection of these two topics is the food industry, specifically the farming industry.

Machine Learning is a somewhat recent, it's been around the 1980s [6]. It has provided many tools and models to identify patterns and solve problems [3]. Relying primarily on data, Machine Learning is a strong candidate to fight this problem of food waste.

There are many drivers which cause food waste along the supply chain, with many of these being partially responsible for food waste. Bad practices come from innapropiate choice of crop varieties, all the way to poor water and nutrient management [2]. Not only farms are responsible, but also supermarkets. As noted by a study done in Brazil, one of the main causes for food waste is "inneffective stock control management" [4]. A report here in the UK also estimates that around 40.7% of the fresh products gets wasted or not completely used [1]. Many scientist belive that effective planning and control over the production can help reduce drastically the amount of wasted food [5].

While it's impossible (or at least very unlikely) to predict weather data to better plan the production of crops, Machine Learning model can provide with an estimate of weather conditions, using data available online. The team is optimistic to building a useful model capable in assisting in the planing phases of crop production.

# 2 Data

The project supervisor was the person responsible for providing the team with the dataset. Given it's sensitivity and confidentiality, this dataset is not publicly. As the main repository is kept public for research purpuses, the dataset will be kept locally. This dataset contains information about crops and weather. Table 1 contains information about each individual sheet on the *.csv* file.

| | |
|---|---|
| Plants | Information about the plants |
| Flight Dates | Information about the flights |
| Planting | Additional information about the plants |
| Weather | Historical weather information |

Table 1: Table with the information per sheet Agrotech

The sheet *Plants* contains our labels to predict, under the column names *Head Weight, Polar Diamter and Radial Diameter*. The type of data in these columns is Float and contain some NULL values. The data consists of 4859 rows. Rows with missing values on the label columns were dropped, as well as rows marked by the *Remove* column (which was also removed).

The *Flight Dates* sheet only contained the Flight Dates for specific Batch Numbers. This sheet was merged with the *Plants* sheet to fill in missing dates. The *Planting* sheet contained additional information about each Batch. It was also merged with the previous sheets. Lastly the *Weather* sheet contains historic weather data.

Features were created from the *Weather* sheet. We used the Plant Date and Flight Date to extract the weather conditions a certain batch was subject to. Some of the features extracted were *Mean Solar Radiation, Total Precipitation, Mean Air Temperature and Humidity.*

The resulting dataset consists of 18 features and 3 labels, with 3236 rows. Most of the features are numerical values, with only one categorical feature. The data was standarised to reduce size difference in dimensions.

# 3 Methodology

With the pre-processing part of the project finished, we now turn towards the actual training of the model. The team used the *ScikitLearn* library in Python to implement the Machine Learning models. The selected models are:

- Linear Regression
- Decision Tree
- Random Forest Regressor
- Gradient Booster Regression

Since this is a Multi-Output Regression problem, these models will be wrapped in a *MultiOutputRegressor* class. This will allow us to build and test the model to predict more than one label at a time. Figure 1 shows the system architecture of the project.

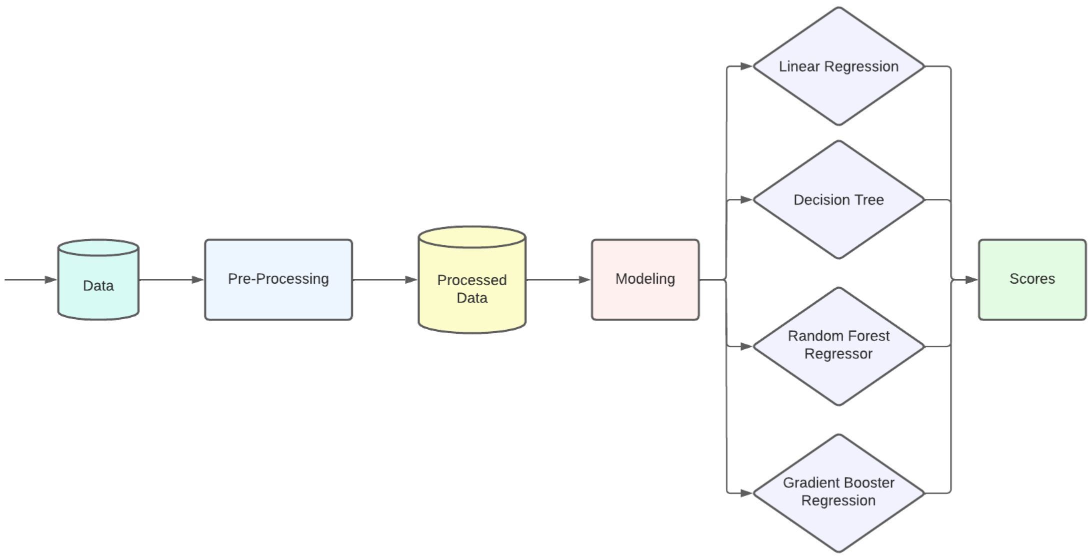


Figure 1: Project Diagram

All the models will be evaluated using 10-fold cross-validation. The metrics to be used for evaluation will be Root Mean Squared Error (RMSE) and $R^2$.

# 4 Results

The results of the training can be seen in Table 2. Figures 2 and 3 also provides comparison between the models developed.

| ML Model | $R^2$ | RMSE |
|---|---|---|
| Linear Regression | 0.68 | 38.03 |
| Decision Tree | 0.64 | 44.36 |
| Random Forest Regression | 0.82 | 32.34 |
| Gradient Booster Regression | 0.82 | 30.73 |

Table 2: Performance per Model

# 5 Discussion

The Linear Regression model achieved an $R^2$ score of 0.68 and an RMSE score of 38.03. Since there are no hyper-parameters in this model no tuning was made.

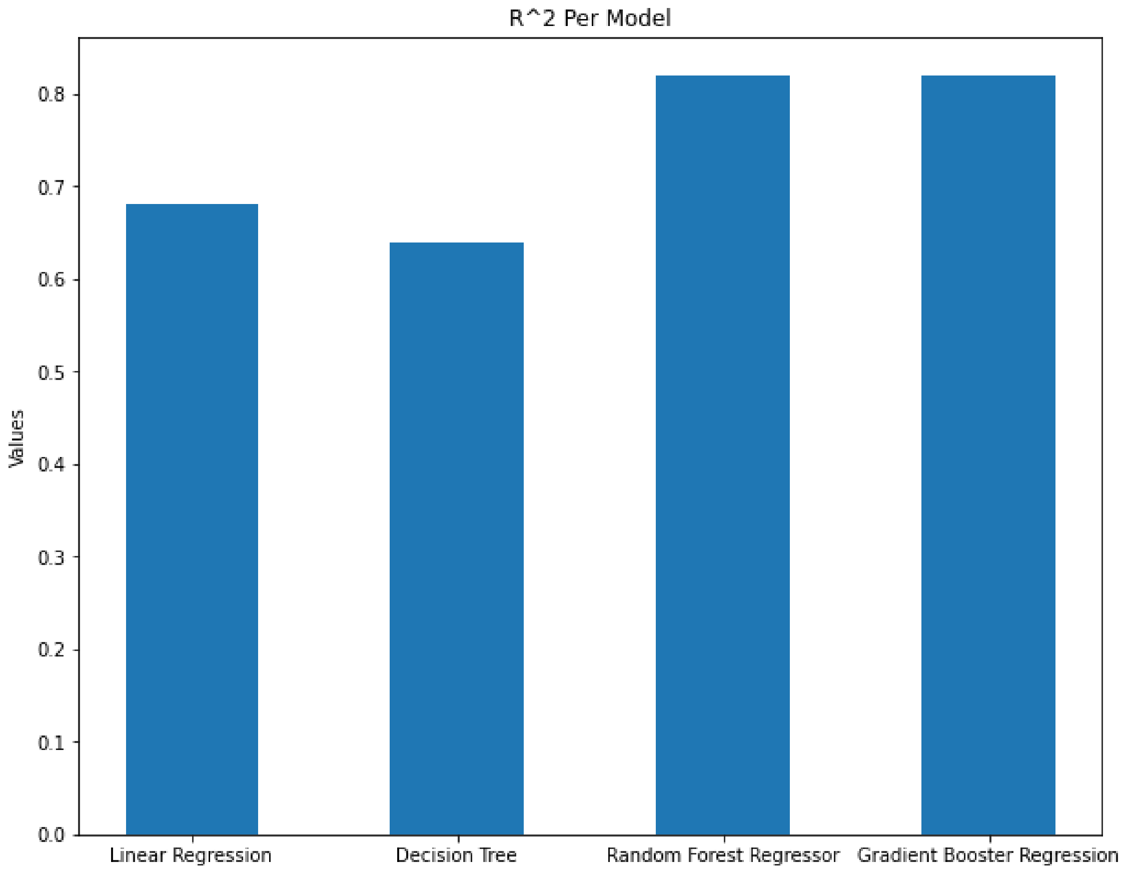


Figure 2: $R^2$ Values per Model

The Decision Tree model achieved an $R^2$ score of 0.64 and an RMSE score of 44.36. The *max_depth* parameter was modified during the training of the model, but no significant improvements came from this tuning.

The Random Forest Regression model achieved an $R^2$ score of 0.82 and an RMSE score of 32.34. The The *n_estimators* parameter was modified in the ranges of (50, 100, 150, 200) with no improvements in the scores but significant increase in the training time the higher the number of estimators, therefore the default value of 100 was used.

The Gradient Booster Regression model achieved an $R^2$ score of 0.82 and as RMSE score of 30.73. The *learning_rate* parameter was tuned but saw no significant improvement on the scores. The *nestimators* parameter was also tuned, but saw no actual improvement on the scores.

Given the data we determied the best model is the Gradient Booster Regression. Figure 2 provides a comparison between the $R^2$ score among the models, while Figure 3 provides a comparison between the RMSE scores among the models.

# 6 Conclusions

The approach used by the team involved using cummulative information from the weather conditions the crop had already been exposed to. This approach doesn't take into account possible future conditions the crops has still yet to face before it's Check Date. While its impossible to accurately predict the weather conditions, historic data from previous years might be useful to improve the performance of the model. This is a potential avenue for future research.

The model built for this project can be used for some rough estimates to help Agrotech better plan the production of their crops and help reduce food waste.

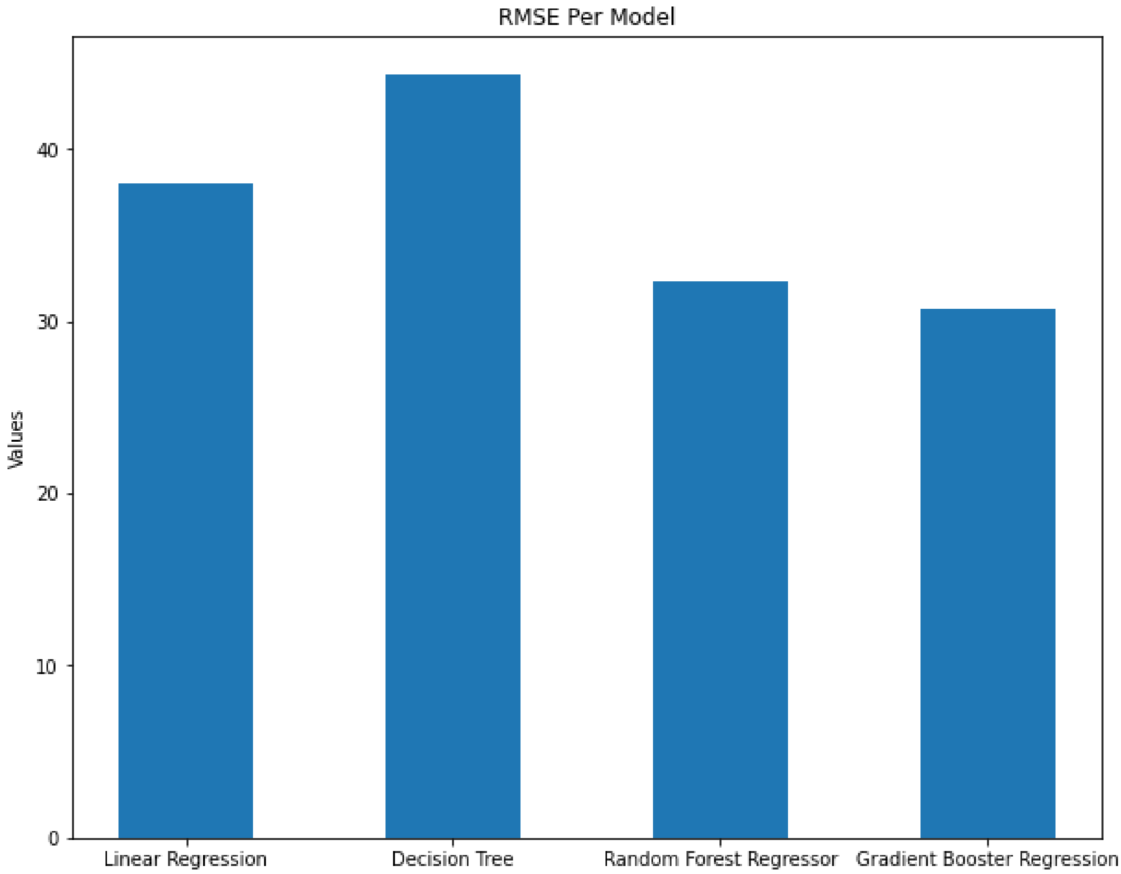


Figure 3: RMSE Values per Model

## References

[1] N. Baker, S. Popay, J. Bennett, and M. Kneafsey. Net yield efficiency: Comparing salad and vegetable waste between community supported agriculture and supermarkets in the uk. *Journal of Agriculture, Food Systems, and Community Development*, 8(4):179, Winter 2019. Copyright - Copyright New Leaf Associates, Inc. Winter/Spring 2019; Last update - 2019-06-19.

[2] M. Blakeney. *Food loss and food waste : causes and solutions.* Edward Elgar Publishing, 2019.

[3] J. E. Brand, B. Koch, J. Xu, P. Atkinson, S. Delamont, A. Cernat, J. W. Sakshaug, and R. A. Williams. *Machine learning.* SAGE Publications Ltd., 2020.

[4] M. W. Eluiza Alberto de, d. N. Caroline Rodrigues, d. F. Michele Gasparoto Moreira Teixeira, and M. V. Mayra. Food waste: an exploratory investigation of causes, practices and consequences perceived by brazilian supermarkets and restaurants. *British Food Journal*, 124(3):1022–1045, 2022. Copyright - © Emerald Publishing Limited 2021; Last update - 2022-02-07; SubjectsTermNotLitGenreText - Denmark; Brazil; Pakistan.

[5] C. M. Galanakis. *Saving Food : Production, Supply Chain, Food Waste and Food Consumption.* Academic Press, an imprint of Elsevier, 2019.

[6] Z.-H. Zhou. *Machine Learning.* Springer Singapore, 2021.